DOSE: Data Selection for Multi-Modal LLMs via Off-the-Shelf Models

Anonymous ACL submission

Abstract

High-quality and diverse multimodal data are essential for improving vision–language models (VLMs), yet existing datasets often contain noisy, redundant, and poorly aligned samples. To address these problems, data filtering is commonly used to enhance the efficiency and performance of multimodal learning, but it introduces extra computational cost because filtering models are usually trained on the same data they are meant to screen. To reduce this cost, we study DOSE, which explores whether off-the-shelf pretrained models that have never seen the target data can be used to select training samples for larger and stronger multimodal models without any task-specific training. Even without fine-tuning, these models can effectively assess text quality and image–text alignment to guide data selection. Based on this, we build a joint quality–alignment distribution and apply adaptive weighted sampling to select informative samples while maintaining long-tail diversity. This approach greatly enhances data diversity and enables models trained on DOSE-filtered data to achieve comparable or even better results than those trained on the full dataset in standard VQA and math benchmarks. Extensive experiments demonstrate the effectiveness, efficiency, and scalability of our method.

1 Introduction

Large Vision-Language Models (LVLMs) have made significant progress in tasks such as image captioning, visual question answering, and instruction following (Zhu et al., 2023; Dai et al., 2023). These models are typically trained in two stages: the first stage performs large-scale image-text pre-training to establish basic vision-language alignment, and the second stage—Visual Instruction Tuning (VIT)—fine-tunes the model on diverse instruction datasets to improve its ability to follow human instructions (Liu et al., 2023b).

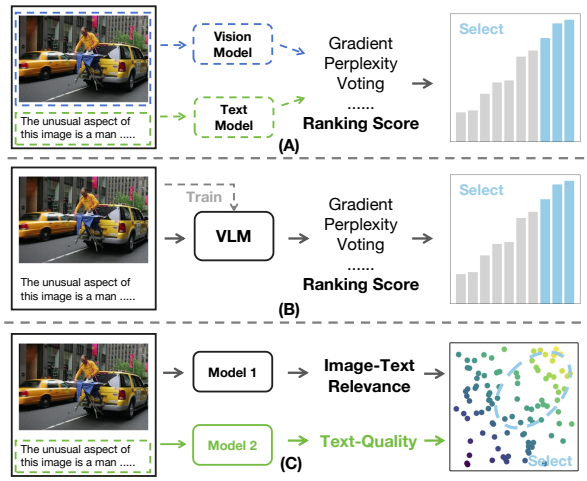


Figure 1: **Comparison of data selection methods. (A)** The methods that rely on a single metric from either vision or text model (dashed line). **(B)** The methods that leverage VLMs for data quality assessment. Notably, the VLMs are already trained on the target data that will be filtered. **(C)** Our approach constructs data distribution by harnessing existing pre-trained models that have not been exposed to the target data.

To enhance task generalization, recent work has expanded the scope of VIT by incorporating a broader range of vision-language tasks. However, training on such large-scale instruction data is computationally expensive and often unaffordable for smaller research labs. Moreover, not all data contribute equally to downstream performance. Prior studies have shown that carefully selected subsets can match or even exceed full-dataset training while significantly reducing cost (Zhou et al., 2023). This motivates a central question: *how can we identify the most valuable data for instruction tuning*?

As shown in Figure 1 (a), existing data selection methods typically rely on proxy metrics derived from the training process, such as loss, perplexity, confidence scores, or gradient norms (Paul

et al., 2021; Chen et al., 2024; Marion et al., 2023a). While these methods are efficient, they often capture only local training dynamics and fail to reflect a sample's semantic richness or generalization utility. They are also tightly coupled with model training, requiring early-stage loss traces or backpropagation, which increases computational overhead and reduces portability, as illustrated in Figure 1 (b) (Cao et al., 2023; Wu et al., 2024; Lee et al., 2024).

In this work, we propose to formulate data quality estimation as a language understanding and reasoning task. The core idea is to leverage the zero-shot semantic and logical capabilities of pretrained large language models (LLMs) (Sachdeva et al., 2024; OpenAI, 2023), and use carefully designed prompts to guide the model to assess the quality of each instruction sample—based on fluency, informativeness, and instruction alignment—purely via forward inference, without relying on any training signals or backpropagation. Compared to traditional proxy-based approaches that depend on loss, confidence, or similar heuristics, LLMs provide a more global, semantically grounded perspective, offering advantages in terms of cost-efficiency, generalizability, and robustness. While vision-language models (VLMs) also possess multimodal capabilities, their training data often overlaps significantly with the target dataset, and their ability to assess linguistic quality is limited. Therefore, we adopt LLMs as a more neutral and reliable evaluator.

However, high-quality scoring alone does not guarantee effective data selection. An equally important component is how samples are chosen based on these scores. To this end, we introduce a lightweight weighted sampling strategy that avoids rigid top-k truncation. Instead of selecting only the highest-ranked examples, our method assigns non-zero sampling probabilities across the score spectrum—preserving rare but informative samples from low-density regions, as illustrated in Figure 1 (c). This helps maintain data diversity, mitigates selection bias, and improves model robustness during training. This LLM-as-evaluator paradigm—combined with a soft, diversity-aware sampling scheme—enables us to construct a compact yet informative data subset for vision-language instruction tuning. By identifying and retaining the most useful examples without overfitting to dominant patterns, our method significantly improves training efficiency while maintaining or even improving final performance.

We conducted extensive evaluations on general VQA benchmarks and specialized math tasks using LLaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2023a) as baselines. Remarkably, DOSE retains 96% of full-data performance on general VQA using only 20% of the data and even surpasses full-data results on math tasks with the same 20% subset. DOSE outperforms methods requiring prior exposure to filtered data, demonstrating superior balance across performance, computational cost, cross-domain generalization, and sample diversity.

Our main contributions are:

- We propose an efficient data selection method that leverages pre-trained, off-the-shelf models to rapidly assess text quality and image-text relevance, significantly reduces the cost of data filtering.
- Extensive experiments demonstrate that our approach achieves an optimal trade-off between selection efficiency and training performance.
- Experiments on multimodal math benchmarks validate that our approach generalizes well to specialized domains, where a small fraction of training data achieves performance comparable to the full training set.

2 Related Work

2.1 Data Quality Scoring

Quality-score was originally developed for importance sampling but is now widely used in training LLMs. The scoring algorithm evaluates sample importance using various methods, including measuring disagreement rates between models (Chitta et al., 2021), assessing whether a sample is likely to be "forgotten" (Toneva et al., 2019), "memorized" (Feldman and Zhang, 2020), or "unlearnable" (Mindermann et al., 2022), and applying perplexity filtering to prioritize low-perplexity samples while discarding high-perplexity ones (Wenzek et al., 2019; Marion et al., 2023b; Muennighoff et al., 2023). Recent advancements have enabled perplexity estimation through efficient model-based simulators, eliminating the need for full LLM inference (Guu et al., 2023). Additionally, some approaches select training data by minimizing the distance between the selected data distribution and high-quality sources such as Wikipedia or books. This is often achieved through contrastive classifiers or feature-space matching (Radford et al., 2019; Anil et al., 2023; Javaheripi et al., 2023). To

more effectively assess the comprehensive quality of multimodal image-text data, we introduce the CLIP-Score (Hessel et al., 2021) for evaluating image-text relevance. For textual data, we leverage the reasoning capabilities of instruction-tuned LLMs to directly evaluate sample quality. Specifically, we use the acceptance probability assigned by the LLM to measure the likelihood that a given text is valid and meaningful.

2.2 Data Selection on Distribution

Data selection is crucial for improving model training quality and can be divided into two categories: distribution-agnostic filtering and distribution-aware selection. Distribution-agnostic methods focus on the quality of individual samples, typically using thresholds to identify subsets. For example, these methods may detect mismatched text-image pairs or misleading elements in images. Specifically, Nguyen et al. (2023); Mahmoud et al. (2023) employ BLIP to identify mismatches between captions and images, while Maini et al. (2023) leverage OCR models to filter images where text is the only feature correlated with the caption. In contrast, distribution-aware methods optimize subset selection by statistically analyzing the overall data distribution. Classical techniques, such as those proposed in Wei et al. (2015); Raskutti and Mahoney (2016); Coleman et al. (2019), aim to maximize subset performance under a fixed budget. More recently, Wang et al. (2023) introduced an approach that replaces traditional models with a trained codebook, clusters samples, and selects representative samples from each cluster. Our method builds upon these ideas by constructing a joint distribution of image-text relevance and text quality. We carefully analyze the impact of different regions and diversity within this joint distribution on data quality, ultimately selecting the most representative samples for training.

3 Methodology

Multimodal data selection mainly focuses on assessment data quality, with existing methods typically assessing text quality and the overall quality of image-text pairs. To achieve comprehensive quality assessment, we combine these methods and create a unified scoring strategy. Existing text quality evaluation methods either introduce bias toward noisy samples with information or face the issue where the evaluation model has already seen the data during training. To address this, we introduce the Text-Quality Score, which leverages the reasoning capabilities of a pre-trained LLM to assess text quality. Additionally, we use the widely adopted CLIP-Score to evaluate the quality of image-text pairs. Meanwhile, selecting data using a static threshold may lead to a loss of diversity and the discarding of valuable edge cases, potentially limiting performance. To address this, we introduce a weighted sampling strategy that integrates data diversity with score-based selection. This approach enables us to select a high-quality subset while maintaining stability and representativeness, ensuring both performance and diversity are preserved.

3.1 Off-the-Shelf Quality Assessment

We leverage the reasoning capabilities of pre-trained LLMs and multimodal language models to evaluate data quality. Inspired by Ask-LLM (Sachdeva et al., 2024), we prompt the LLM to predict whether an input sample is suitable for fine-tuning a multimodal language model. As illustrated in Table 3, the LLM predicts "yes" when the text is informative, well-formatted, and aligned with visual instruction tuning objectives. The softmax probability assigned to the "yes" token serves as the *Text-Quality Score* for the sample.

In addition, similar to (Nguyen et al., 2023; Mahmoud et al., 2023; Maini et al., 2023; Fang et al., 2023), we use the CLIP-ViT-B32 (OpenAI, 2023) to obtain CLIP-Score (Hessel et al., 2021) to assess the alignment between images and their captions. The CLIP model projects both images and text into a shared embedding space, and the cosine similarity between these embeddings quantitatively measures the image-text relevance.

3.2 Weighted Random Sampling

To effectively *select high-quality and diverse samples*, we first construct a target distribution $q(x)$ based on the original score distribution $p(x)$. This new distribution shifts the density toward high-quality regions while increasing the probability of sampling rare but informative long-tail examples, mitigating the over-representation of moderate-quality, high-density regions. We then perform *Weighted Random Sampling (WRS)* based on $q(x)$, assigning higher sampling probabilities to desirable samples. This strategy enables a balanced selection that promotes both sample quality and diversity, with the detailed filtering procedure and corresponding results illustrated in the Appendix

(see Figure 5).

Sampling Procedure We begin by computing the statistical properties of the score distribution, including the mean μ_{data} and standard deviation σ_{data}. To obtain a smooth estimate of the data distribution, we apply Kernel Density Estimation (KDE):

$$KDE(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right), \quad (1)$$

where $K(\cdot)$ is the Gaussian kernel, N is the number of samples, and h is the bandwidth.We then apply KDE to the filtered data and identify the principal mode of the distribution:

$$\mu_{\text{peak_kde}} = \arg \max_{x \in [x_{\min}, x_{\max}]} KDE(x). \quad (2)$$

Next, we get the maximum score as:

$$x_{max} = \max x_i, \quad (3)$$

To determine a robust target center, we combine two complementary indicators: the KDE mode $\mu_{\text{peak_kde}}$, which captures the most representative high-density region of the distribution, and x_{max}, which reflects the highest-quality sample not considered an outlier. Averaging these two values balances the typical quality level with the upper bound of acceptable sample quality, while avoiding the skew often introduced by means or medians in imbalanced data. The final target center is defined as:

$$\mu_{\text{peak_wrs}} = \frac{\mu_{\text{peak_kde}} + x_{max}}{2}. \quad (4)$$

Based on $\mu_{\text{peak_wrs}}$, we define the target distribution $q(x)$ and the original distribution $p(x)$ as Gaussian distributions centered at $\mu_{\text{peak_wrs}}$ and μ_p, respectively. Their corresponding probability density functions are given as follows:

$$\begin{aligned} q(x) &= \mathcal{N}\big(x;\, \mu_{peak_wrs},\, \sigma_{\text{data}}\big), \\ p(x) &= \mathcal{N}\big(x;\, \mu_{peak_kde},\, \sigma_{\text{data}}\big). \end{aligned} \quad (5)$$

where μ_{peak} is the mean of the target distribution, and σ_{data} is the standard deviation (consistent with the original data). To perform WRS, we calculate the weight for each data point x_i as the ratio of the probability density under the target distribution to that under the original distribution:

$$w_i = \frac{q(x_i)}{p(x_i) + \epsilon}, \quad (6)$$

where $\epsilon = 10^{-10}$ is a small constant added to avoid division by zero. Subsequently, we normalize the weights:

$$w_i' = \frac{w_i}{\sum_{j=1}^{N} w_j}. \quad (7)$$

Finally, based on the normalized weights w_i', we perform weighted random sampling to select M samples (without replacement) from the original data:

$$S_x = \{x_{i_1}, x_{i_2}, \ldots, x_{i_M}\}, \quad (8)$$

where i_k are indices randomly drawn according to the weights w_i'. Through these steps, we generate a new sample set S that better aligns with the characteristics of the target distribution $q(x)$. Also, based on the Image-Text Relevance Scores (y_i), we can apply the same sampling strategy to obtain the corresponding subset:

$$S_y = \{y_{i_1}, y_{i_2}, \ldots, y_{i_M}\}, \quad (9)$$

Combined Sampling Once the positions of all data points are determined in a two-dimensional coordinate space—where each point is defined by x_i (text quality) and y_i (image-text relevance)—we construct a density-like distribution that captures the frequency of data points within local regions. This distribution reveals patterns in the data, enabling us to analyze and compare the data distribution before and after sampling. Based on this distribution, we design a sampling strategy that prioritizes regions with both high densities and favorable characteristics in terms of x_i and y_i. Specifically, we define subsets S_x and S_y, which capture key features along the x_i and y_i dimensions, respectively. By combining the intersection of S_x and S_y, we derive the final sampling results.

$$\textit{DOSE} = \{(x_i, y_i) \mid (x_i, y_i) \in S_x \cap S_y\}. \quad (10)$$

This approach ensures that the sampled points not only reflect the underlying data distribution but also align with preferred ranges for text quality and image-text relevance.

4 Experiments

We evaluate the effectiveness of our proposed data selection method, DOSE, on a range of visual instruction tuning (VIT) tasks. Our goal is to assess whether DOSE can identify high-value training subsets that achieve competitive or superior performance to baseline methods under a fixed data budget. This section describes our implementation, experimental setup, comparisons against baselines, and efficiency analysis.

Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	MME	MMBench en	MMBench cn	LLaVA-W Bench	Rel. (%)
Full	79.1	63.0	47.8	68.4	58.2	1476.9	66.1	58.9	67.9	100
Methods that already used full data before data selection										
COINCIDE	76.5	59.8	46.8	69.2	55.6	1495.6	63.1	54.5	67.3	97.4
ICONS	76.3	60.7	50.1	70.8	55.6	1485.7	63.1	55.8	66.1	98.6
Methods that never used full data before data selection										
Random	75.7	57.6	44.7	66.5	54.2	1389.0	62.2	54.8	65.0	94.5
CLIP-Score	73.4	51.4	43.0	65.0	**54.7**	1331.6	55.2	52.0	66.2	91.2
EL2N	76.2	58.7	43.7	65.5	53.0	1439.5	53.2	47.4	64.9	92.0
Perplexity	75.8	57.0	47.8	65.1	52.8	1341.4	52.0	45.8	68.3	91.6
SemDeDup	74.2	54.5	46.9	65.8	55.5	1376.9	52.2	48.5	**70.0**	92.6
D2-Pruning	73.0	58.4	41.9	**69.3**	51.8	1391.2	**65.7**	**57.6**	63.9	94.8
Self-Sup	74.9	59.5	46.0	67.8	49.3	1335.9	61.4	53.8	63.3	93.4
Self-Filter	73.7	58.3	**53.2**	61.4	52.9	1306.2	48.8	45.3	64.9	90.9
Ours	**77.3**	**58.7**	46.5	67.2	54.4	**1462.2**	62.5	54.8	65.8	**96.0**

Table 1: **Comparisons with baseline methods.** For a fair comparison, all models are trained by 20% of full training data and the data subsets are selected by different methods. The best results among methods that do not access the full training data before selection are shown in **bold**.

Implementation Details All experiments are conducted using the LLaVA-1.5 architecture, focusing on the second-stage VIT process where multimodal instruction-following models are trained on image–text–instruction triples. For the main experiments, we adopt the 7B version of LLaVA-1.5 and apply DOSE to select a 20% subset (665K samples) from the official VIT dataset, while keeping the pretrained vision and language encoders frozen and retraining only the instruction-tuning stage. Each training sample is assigned two quality scores: a text quality score predicted by Vicuna-7B (Team, 2023) and an image–text alignment score computed with CLIP-Score (Hessel et al., 2021). These scores define a joint 2D distribution in which each point represents a sample's textual and visual informativeness (as illustrated in Figure 5 in the Appendix). We then perform Weighted Reservoir Sampling (WRS) over this 2D space to construct the final training subset, prioritizing samples that are strong in both modalities. To further assess downstream generalization, we also apply DOSE to the MathV360k dataset (Shi et al., 2024) and fine-tune LLaVA-1.5-13B; this experiment serves as a case study and is not included in leaderboard comparisons.

Experimental Setup We evaluate DOSE on nine diverse VIT benchmarks: VQAv2, GQA, VizWiz, SQA-I, TextVQA, POPE, MME, MMBench (English), and LLaVA-W. Complete dataset details are provided in the Appendix. Each model is trained using only 20% of the available VIT data. We report absolute scores using each benchmark's official evaluation metric, and compute Relative Performance (Rel.%) by normalizing against the score achieved using the full 100% VIT training set. To contextualize these results, we also compare DOSE against two categories of data selection methods:

Seen-data selectors, such as ICONS (Wu et al., 2024) and COINCIDE (Lee et al., 2024), which rely on full-data finetuning to score or cluster training samples.

Unseen-data selectors, which operate without accessing the full dataset and include Random sampling, CLIP-Score, EL2N (Paul et al., 2021), Perplexity (Marion et al., 2023a), SemDeDup (Abbas et al., 2023), D2-Pruning (Maharana et al., 2023), Self-Sup (Sorscher et al., 2022), and Self-Filter (Chen et al., 2024).

4.1 Main Results and Analysis

As shown in Table 2, DOSE achieves the highest average relative performance (96.0%) among all unseen-data baselines. It improves upon D2-Pruning (94.8%) and Self-Filter (93.2%), and reduces the performance gap to full-data methods such as ICONS (98.6%) and COINCIDE (97.4%) to less than 3 percentage points. DOSE consistently outperforms Random across all nine benchmarks (e.g., GQA: 58.6 vs 57.6; TextVQA: 54.4 vs 54.2), and is competitive with or better than more complex selection methods.

While DOSE does not match the top-line scores of full-data selectors, this is expected: ICONS and COINCIDE rely on full-data supervision and ex-

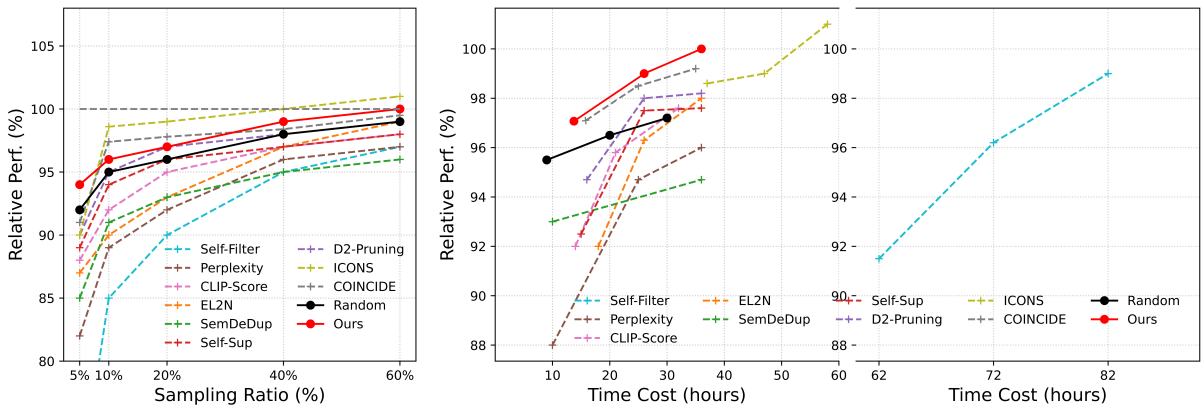


Figure 2: DOSE Data-Selection Efficiency and Wall-Clock Time Trade-Offs. (Left) Average relative performances of all coreset selection techniques at different sampling ratios for the LLaVA-1.5 dataset. (Right) Comparison of coreset selection techniques on average relative performance and wall-clock time cost. The wall-clock time cost includes both the data selection and finetuning of the target VLM. The time cost is measured in hours of running time on a computing node with 4×V100 GPUs. The left panel presents the average relative performance across sampling ratios of 20%, 40%, and 60%.

ploit task-specific model feedback. In contrast, DOSE requires no additional training, no full-dataset traversal, and operates entirely on pre-trained model scores. This leads to significantly lower computational cost and stronger generality, making it particularly suitable for scalable or resource-limited settings.

Different Selection Ratio. As shown in Figure 2 (Left), we compare DOSE (red solid line with circles) against ten baselines—Random (black), Perplexity (Marion et al., 2023a), CLIP-Score (Hessel et al., 2021), EL2N (Paul et al., 2021), SemDeDup (Abbas et al., 2023), Self-Sup (Sorscher et al., 2022), D2-Pruning (Maharana et al., 2023), COINCIDE (Lee et al., 2024), ICONS (Wu et al., 2024), and Self-Filter—across sampling ratios from 5 % to 60 %. DOSE rapidly climbs to 99 % Rel. by 40 % sampling, matching or exceeding all other unseen-data methods and even approaching the seen-data ICONS (Wu et al., 2024) curve at higher ratios.

Efficiency and Performance. Among all data selection baselines showen in Figure 2 (Right), DOSE achieves the largest performance gains among methods that do not rely on prior exposure to the training data, outperforming baselines such as Random, CLIP-Score, EL2N, SemDeDup, Perplexity, Self-Sup, D2-Pruning, and Self-Filter by 1–4 percentage points under identical sampling ratios and time budgets. Even against the two leading seen-data methods, ICONS and COINCIDE, DOSE holds clear advantages. ICONS and COINCIDE both require an expensive full-data fine-tuning pass before sample selection—a cost that would recur for any new dataset yet is omitted from their reported compute comparisons—whereas DOSE skips this phase entirely, relying solely on off-the-shelf pre-trained models for scoring and weighted sampling. As a result, direct comparisons of compute costs are misleading. Moreover, DOSE's linear-time scoring lets it reach 97.4 % relative performance in 12 h and 98.5 % in 22 h, whereas COINCIDE needs 15 h/97.4 % and 25 h/98.4 %, and ICONS—lacking a time-optimized pipeline—lags further behind. Finally, DOSE requires no clustering hyperparameters, gradient-influence computations, or extra network training—its runtime scales linearly with dataset size and is immediately deployable—while seen-data methods add complexity that complicates tuning and extension.

4.2 Unseen-task Generalization.

As shown in Table 2, we filtered the MathV360K (Shi et al., 2024) dataset and performed continuous fine-tuning on LLaVA-1.5-13B (Liu et al., 2023a) using high-quality subsets of varying proportions. In this process, we strictly adhered to the experimental settings of Math-LLaVA (Shi et al., 2024). Since the evaluation on MathVista requires GPT-3.5 (Brown et al., 2020) to extract key results, and the performance of different period versions may vary, we reproduced the results of Math-LLaVA as a benchmark for comparison. The experimental results demonstrate that our method achieves performance comparable to Math-LLaVA (Shi et al., 2024) when using only 20% of the high-quality data. Furthermore, when

Size	Math-LLaVA on MathVista													
	FQA	GPS	MWP	TQA	VQA	ALG	ARI	GEO	LOG	NUM	SCI	STA	Rel.%	Aver.
Random selection on MathV360K														
5%	22.7	38.0	30.7	41.1	38.6	36.7	31.4	38.1	21.6	30.6	38.5	23.9	88.4	32.7
20%	30.9	44.2	42.9	39.9	33.5	39.9	36.5	43.9	28.8	27.8	45.1	29.6	98.7	36.9
40%	32.3	52.4	43.0	37.3	35.2	45.6	35.7	52.3	16.2	27.8	41.9	35.9	97.6	38.0
DOSE selection on MathV360K														
5%	33.4	38.9	30.1	36.1	34.1	36.3	29.5	36.8	24.3	26.4	36.1	31.9	88.4	32.8
10%	30.5	39.9	33.9	39.9	31.8	37.4	30.0	40.2	16.2	26.7	40.2	31.9	86.8	33.2
20%	33.1	45.7	45.7	42.4	**36.9**	43.1	38.5	45.2	**29.7**	**31.3**	41.0	35.9	**104.8**	39.1
40%	32.7	49.5	47.3	43.7	34.6	47.0	37.1	49.4	18.9	27.8	40.2	37.5	100.4	38.8
65%	30.5	49.5	**53.8**	42.4	29.1	44.8	37.4	48.5	8.1	24.3	41.9	37.5	93.1	37.3
80%	32.4	**53.4**	49.5	**45.6**	36.3	**48.4**	**39.4**	**51.9**	16.2	27.8	**46.7**	38.2	103.5	**40.5**
100%†	**37.9**	52.8	46.8	44.3	27.9	**48.4**	33.2	**51.9**	18.9	23.6	45.1	**41.9**	100	39.4

Table 2: **Comparison with different data selection scales on domain-specific benchmarks.** † represents our reproduced results of Math-LLaVA-13B. The best results in all tasks are in bold. MathVista is divided in two ways: task type or mathematical skill, and we report the accuracy under each subset. Rel.% keep same setting with general benchmarks, and Aver. means the average score of all tasks.

using 80% of the data, the overall performance of the model improves by 1 percentage point. While DOSE generally performs well, it occasionally underperforms compared to random sampling on tasks like GPS (40%), TQA (5%), and VQA (5%) under small sampling ratios. This is mainly because high-score samples tend to cluster in semantically similar regions, leading to reduced diversity and limited generalization. In contrast, random sampling retains a broader variety of examples, which can be more effective in certain tasks.

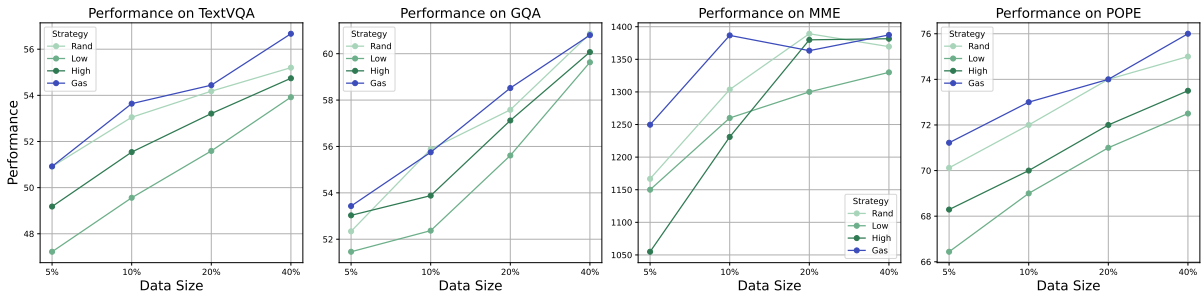
4.3 Ablation Study

In this section, we conduct ablation experiments by comparing different scoring strategies, score-based sampling strategies, and the fusion of these two strategies. The results are presented in Figure 3a, Figure 3b, and Figure 4 in Appendix.

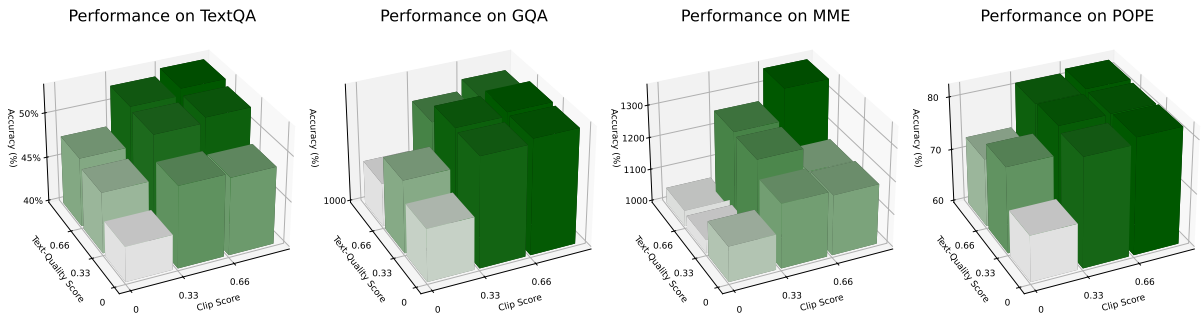
Effectiveness of Filtering Methods To evaluate the effectiveness of Text-Quality and CLIP scores independently, we conduct controlled experiments in Stage 2 of the LLaVA training pipeline, as shown in Figure 3a. We compare four sampling strategies using the Text-Quality Score: *Rand* (random sampling), *High* (top-scoring filtering), *Low* (low-scoring filtering), and *Gas* (Gaussian-based weighted random sampling that balances quality and diversity). Overall, the *High* strategy outperforms *Low*, demonstrating the validity of the Text-Quality score in assessing data quality. We further observe in Figure 3a that at a 40% sampling ratio, *Rand* surpasses *High* on several benchmarks. As discussed in Section 4.3, score-based filtering tends to concentrate on samples with similar language and structure, reducing task diversity and generalization. In contrast, random sampling naturally preserves variation in task types and styles, sometimes yielding better performance. These findings highlight the motivation behind our proposed DOSE framework: combining quality-driven scoring with diversity-aware sampling to achieve a better trade-off. The *Gas* strategy, which embodies this principle, consistently outperforms *Rand*, confirming the effectiveness of our data selection method.

In our evaluation of image-text relevance, shown in Figure 4, we compared four sampling strategies using the CLIP Score. The results revealed that the "Gas" strategy significantly outperformed the others. This suggests that as the filtering ratio decreases, data quality differences become more noticeable, making it suitable for large datasets with low usage needs. However, as the dataset size grows, the differences in quality between filtered and unfiltered data become smaller. We also found that in the GQA task, the data filtered by CLIP Score did not show significant advantages, likely because the original data already had strong image-text relevance. Such findings underscore the rationale for DOSE, which unifies quality-based and diversity-aware signals to overcome the weaknesses of single-score filtering.

Effectiveness of Combined Sampling As shown in Figure 3b, we identified 9 candidate regions



(a) Performance comparison of different strategies based on Text-Quality Score on TextVQA, GQA, MME, and POPE datasets.



(b) Performance comparison of different regions in the combined distribution, based on CLIP-Score as X-axis and Text-Quality Score as Y-axis.

Figure 3: Overall performance comparisons across different strategies and datasets. (a) and (b) shows the results of the ablation study on the combined distribution, where the height of the columns indicates that taller columns correspond to a darker shade of green.

based on the original data distribution. These regions represent clusters of data, reflecting the similarities and differences among samples. To create the combined distribution sampling data, we randomly sampled 5% of the overall data from each candidate region. This method ensures diversity in the samples while effectively capturing the underlying structure of the data. After constructing the combined distribution sampling data, we trained the model using the same settings as the single-method approach and tested it on several datasets, including TextQA (Singh et al., 2019a), GQA (Hudson and Manning, 2019a), POPE (Li et al., 2023a), and MME (Fu et al., 2023). And, the performance results are shown in Figure 3b, which indicate that in the upper right area—where both CLIP and Text-Quality Score are high—the model generally performs better. This suggests that in general task, the combination of the two sampling methods can effectively select data that helps improve the model's performance. By using this combined sampling method based on the distribution, we enhance the representativeness and quality of the data, thereby improving the model's training efficiency.

5 Conclusion

In this work, we propose DOSE, an efficient and practical data selection method for multimodal instruction tuning. DOSE leverages off-the-shelf models to evaluate text quality and image-text alignment separately, then combines these scores into a unified quality-alignment distribution for adaptive weighted random sampling. This approach preserves data diversity while identifying the most informative samples. Our experimental evaluation demonstrates DOSE's effectiveness across multiple dimensions. On both general VQA tasks and specialized math benchmarks, DOSE achieves comparable performance to full-dataset training using only 20% of the data, and surpasses full-dataset results when using 40% to 80% subsets. Crucially, DOSE outperforms existing unseen-data selection strategies in both effectiveness and computational efficiency, while operating entirely at inference time without requiring fine-tuning or additional training. These results underscore the critical importance of high-quality data selection in multimodal learning and establish DOSE as a scalable, practical solution for resource-constrained environments.

6 Limitations

While our method demonstrates strong performance and high efficiency, our study is constrained by the experimental cost and a limited exploration budget. We evaluated only an array of sampling ratios and primarily tested our method on LLaVA-1.5 models (7B & 13B), without assessing more fine-grained sampling ratios or more types of models. As a result, the generality of DOSE across additional sampling ratios and diverse architectures remains to be validated in future work.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, and Zhifeng Chen et al. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Liangliang Cao, Bowen Zhang, Chen Chen, Yinfei Yang, Xianzhi Du, Wencong Zhang, Zhiyun Lu, and Yantao Zheng. 2023. Less is more: Removing text-regions improves clip training efficiency and robustness. *arXiv preprint arXiv:2305.05095*.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. 2024. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. *arXiv preprint arXiv:2402.12501*.

Kashyap Chitta, José M Álvarez, Elmar Haussmann, and Clément Farabet. 2021. Training data subset search with ensemble active learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):14741–14752.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. 2019. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Kelvin Guu, Albert Webson, Ellie Pavlick, Lucas Dixon, Ian Tenney, and Tolga Bolukbasi. 2023. Simfluence: Modeling the influence of individual training examples by simulating training runs. *arXiv preprint arXiv:2303.08114*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Drew A Hudson and Christopher D Manning. 2019a. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Drew A Hudson and Christopher D Manning. 2019b. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, and 1 others. 2023. Phi-2: The surprising power of small language models.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. 2024. Concept-skill transferability-based data selection for large vision-language models. *arXiv preprint arXiv:2406.10995*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *CoRR*, abs/2310.02255.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. 2023. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. *arXiv preprint arXiv:2310.07931*.

Anas Mahmoud, Mostafa Elhoushi, Amro Abbas, Yu Yang, Newsha Ardalani, Hugh Leather, and Ari Morcos. 2023. Sieve: Multimodal dataset pruning using image captioning models. *arXiv preprint arXiv:2310.02110*.

Pratyush Maini, Sachin Goyal, Zachary C Lipton, J Zico Kolter, and Aditi Raghunathan. 2023. T-mars: Improving visual representations by circumventing text feature learning. *arXiv preprint arXiv:2307.03132*.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023a. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023b. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and 1 others. 2022. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pages 15630–15649. PMLR.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36:22047–22069.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Garvesh Raskutti and Michael W Mahoney. 2016. A statistical perspective on randomized sketching for ordinary least-squares. *The Journal of Machine Learning Research*, 17(1):7508–7538.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. 2024. How to train data-efficient llms. *arXiv preprint arXiv:2402.09668*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models.

In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680, Miami, Florida, USA. Association for Computational Linguistics.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019a. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019b. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536.

The Vicuna Team. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna.

M. Toneva, A. Sordoni, R. Combes, A. Trischler, Y. Bengio, and G. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *ICLR*.

Alex Jinpeng Wang, Kevin Qinghong Lin, David Junhao Zhang, Stan Weixian Lei, and Mike Zheng Shou. 2023. Too large; data reduction for vision-language pre-training. *arXiv preprint arXiv:2305.20087*.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. 2015. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pages 1954–1963. PMLR.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. 2024. Icons: Influence consensus for vision-language data selection. *arXiv preprint arXiv:2501.00654*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A Filtering Details

We applied the DBSCAN algorithm to filter out anomalous noise from the candidate data. This subset accounted for approximately 0.01% of the entire dataset, corresponding to a total of 63 samples.

B Benchmarks

GQA (Hudson and Manning, 2019b), which focuses on reasoning about visual attributes like color and shape, and VQA-v2 (Goyal et al., 2017), which assesses broader visual reasoning. MME (Fu et al., 2024) evaluates both perceptual abilities and cognitive reasoning, while TextVQA (Singh et al., 2019b) tests OCR-based reasoning. POPE (Li et al., 2023b) addresses object hallucination, assessing models' ability to avoid generating non-existent objects. VizWiz (Gurari et al., 2018) focuses on basic visual reasoning for users who are blind, and ScienceQA (Lu et al., 2022) evaluates knowledge-grounded question answering. Together, these benchmarks provide a comprehensive test of reasoning, perception, and understanding. Meanwhile, for the Special VQA task, we use MathVista (Lu et al., 2023), a benchmark designed to assess mathematical reasoning in visual contexts. It comprises 6,141 questions from various datasets and covers categories such as FQA, GPS, MWP, TQA, and VQA. With a focus on arithmetic, algebra, and logic, MathVista includes a diverse range of image types, making it an essential platform for evaluating models' capabilities in mathematical reasoning.

C Result Analysis

To understand how our proposed data selection strategy enhances training performance and efficiency, we conducted a visualization and analysis of the data used in LLaVA stage 2, consisting of 665k data points. In the left panel of Figure 5, we plotted the CLIP-Score and Text-Quality Score for each data point, revealing a significant concentration of data points in the central area. This suggests that the data likely follows a normal distribution in both scores, indicating regions of higher data quality. These insights led us to examine performance variations across different regions, as discussed in Section 4.3. We found that areas with higher concentrations of data points generally correlated with better performance. This understanding drove us to combine these insights with WRS to create a high-quality data subset selection strategy.

We then visualized the distributions resulting from random sampling (light blue) and WRS sampling (light green) in the right panel of Figure 5. The WRS sampling distribution shows a pronounced concentration in regions with higher CLIP and Text-Quality Scores, effectively validating our strategy for assessing data quality and demonstrating the benefits of our sampling approach.

D Time Cost Analysis

Figure 2 presents a joint analysis of model performance and wall-clock cost across different data selection strategies. The left panel reports the average relative performance of each method under varying sampling ratios (20%, 40%, 60%), while the right panel compares the corresponding total wall-clock time, including both data selection and fine-tuning. Each curve comprises three data points representing these ratios.

Although the x-axes differ (sampling ratio vs. total time), the relationship is direct—higher sampling ratios typically incur greater computational cost. This visualization highlights how different methods navigate the trade-off between efficiency and effectiveness.

Among the methods, Perplexity-based filtering exhibits the steepest increase in time cost as the sampling ratio grows. This is due to its inherently sequential and non-parallelizable scoring process, which requires token-level log-likelihood computation for every instruction–response pair. Additionally, Perplexity re-evaluates the selected samples from scratch at each ratio, leading to near-linear or worse scaling behavior in wall-clock time. This limits its scalability to large datasets.

Consistent with prior works such as ICONS (Wu et al., 2024) and COINCIDE (Lee et al., 2024), we omit the full-data training cost–performance curve in this figure, as the focus is on fixed-ratio comparisons to highlight efficiency gains.

Tasks	Examples of Task Templates
Original Template	**Question**: " $\langle image \rangle$ What are the colors of the bus in the image? " **Answer**: " The bus in the image is white and red. "
Scoring Template	**Question**: " ### What are the colors of the bus in the image? The bus in the image is white and red. ### Does the previous paragraph demarcated within ### contain informative signal for visual instruction tuning a vision-language model? An informative data point should be well-formatted, contain usable knowledge of the world, and strictly NOT have any harmful, racist, sexist, etc. content. OPTIONS: -yes -no " **Answer**: " Response: yes"

Table 3: Task template examples. "Original Template" represents the original format of the data, while "Scoring Template" represents the format used to assist in evaluating the quality of the text within the data. $\langle image \rangle$ indicates that the original data contains corresponding image information; in the scoring template, we only assess the quality of the textual information, so this token is omitted.

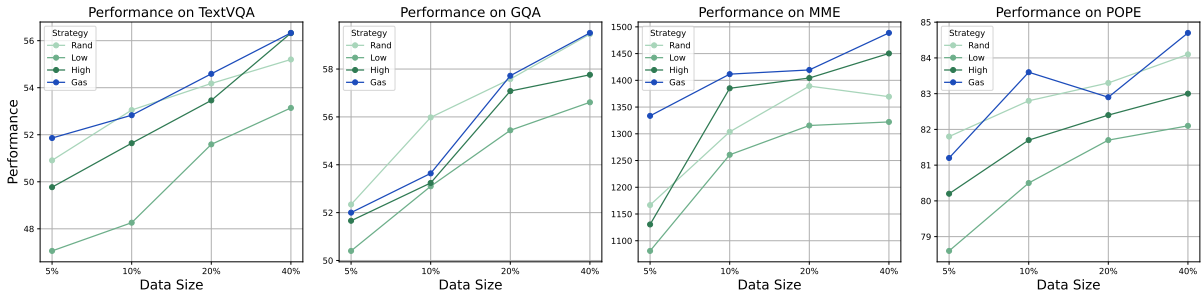


Figure 4: Performance comparison of different strategies based on CLIP-Score on TextVQA, GQA, MME, and POPE datasets.

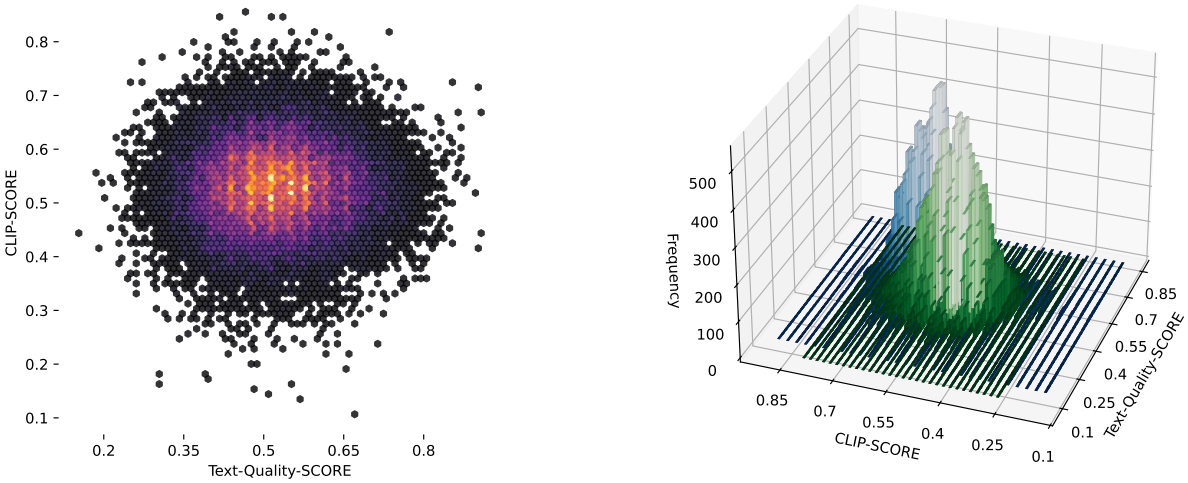


Figure 5: **(Left) The combined distribution of Text-Quality and CLIP Score.** The combined distribution is plotted with Text-Quality Score on the X-axis and CLIP Score on the Y-axis, forming a 2D distribution. The density is illustrated, where lighter colors indicate lower densities and brighter colors represent higher densities. **(Right) The combined distribution of sampling results of 665K data of LLaVA Stage 2.** The same axis settings as the left figure are used, with an additional z-axis representing the data density. The height of the z-axis corresponds to the density of data in the respective region.